High-fidelity synthetic patient data applications and privacy considerations

Received: 8th May, 2024

Puja Myles

Director, Clinical Practice Research Datalink, Medicines and Healthcare Products Regulatory Agency, UK

Puja Myles is Director of the Medicines and Healthcare Products Regulatory Agency's (MHRA) specialist real world data research services centre, Clinical Practice Research Datalink (CPRD). She initially joined the MHRA as Head of Observational Research, CPRD in 2017 and, prior to this, trained as a public health specialist and was a public health academic at the University of Nottingham, UK. She is a fellow of the Faculty of Public Health (UK), a senior fellow of the Higher Education Academy (UK) and has a doctorate in epidemiology. She has been the MHRA's strategic lead on the development of synthetic data generation approaches and applications since 2017.

MHRA, 10 South Colonnade, Canary Wharf, London, E14 4PU, UK
E-mail: puja.myles@mhra.gov.uk

Colin Mitchell

Head of Humanities, PHG Foundation, University of Cambridge, UK

Colin Mitchell is Head of Humanities at the PHG Foundation, a health policy unit and part of the University of Cambridge. The PHG Foundation's multidisciplinary team works with health professionals, researchers and policymakers to explore the implications of emerging data and related technologies for health care and research. Colin leads the Foundation's work on legal and ethical issues arising from novel health technologies, biomedical research and data-driven innovation. Colin has a PhD in health law from the University of Amsterdam, a master's of studies in legal research from the University of Oxford and a BA in law from the University of Cambridge.

PHG Foundation, 2 Worts Causeway, Cambridge, CB1 8RN, UK
E-mail: colin.mitchell@phgfoundation.org

Elizabeth Redrup Hill

Senior Policy Analyst (Law and Regulation), PHG Foundation, UK

Elizabeth Redrup Hill is a senior policy analyst (law and regulation) in the Humanities Team at the PHG Foundation, a health policy unit and part of the University of Cambridge. The PHG Foundation's multidisciplinary team works with health professionals, researchers and policymakers to explore the implications of emerging data and related technologies for healthcare and research. Elizabeth holds a doctorate (PhD) in medical law and ethics, and an LLB in law from the University of Southampton, UK. She is also an external lecturer in medical law and ethics at Imperial College London. Her research interests are in the regulatory and ethical challenges for genomic and health data, artificial intelligence and other innovative health technologies.

PHG Foundation, 2 Worts Causeway, Cambridge, CB1 8RN, UK
E-mail: elizabeth.redrup@phgfoundation.org

Luca Foschini

President, Sage Bionetworks, USA

Dr Foschini's research in the past decade has focused on the emerging field of digital medicine, particularly in the areas of data collection and analysis methodology, with topics including machine learning in health care, continuous health monitoring and privacy in high-dimensional data. Dr Foschini holds a PhD in computer science from the University of California, Santa Barbara and a master's degree in computer engineering from the Sant'Anna School of Pisa. He has also conducted theoretical

computer science and cybersecurity research in academia and industry, including research positions at CERN and Google.

Sage Bionetworks, 2901 Third Ave. Suite 330, Seattle, WA 98121, USA
E-mail: luca.foschini@sagebionetworks.org

Zhenchen Wang

Head of Data Analytics and Machine Learning, Scientific Data and Insights, Medicines and Healthcare products Regulatory Agency (MHRA), UK

Zhenchen Wang is an innovator in the intersection of artificial intelligence (AI) and machine learning (ML) within healthcare data applications. With a solid foundation in computer science and a deep understanding of electronic healthcare systems, Zhenchen has dedicated his career to revolutionising the way healthcare data is analysed and utilised to improve patient safety and operational efficiency. He joined MHRA in 2017 and has been the technical lead on research and development of synthetic data generation. Zhenchen leads a team of talented data scientists and engineers in developing cutting-edge AI and ML algorithms tailored specifically for healthcare applications. His expertise lies in leveraging advanced analytics techniques to extract valuable insights from complex medical datasets.

MHRA, 10 South Colonnade, Canary Wharf, London, E14 4PU, UK
E-mail: zhenchen.wang@mhra.gov.uk

Abstract This paper explores the potential applications of high-fidelity synthetic patient data in the context of healthcare research, including challenges and benefits. The paper starts by defining synthetic data, types of synthetic data and approaches to generating synthetic data. It then discusses the potential applications of synthetic data in addition to as a privacy enhancing technology and current debates around whether synthetic data should be considered personal data and,therefore, should be subjected to privacy controls to minimise reidentification risks. This will be followed by a discussion of privacy preservation approaches and privacy metrics that can be applied in the context of synthetic data. The paper includes a case study based on synthetic electronic healthcare record data from the Clinical Practice Research Datalink on how privacy concerns due to reidentification have been addressed in order to make this data available for research purposes. The authors conclude that synthetic data, particularly high-fidelity synthetic patient data, has the potential to add value over and above real data for public health and that it is possible to address privacy concerns to make synthetic data available via a combination of privacy measures applied during the synthetic data generation process and post-generation reidentification risk assessments as part of data protection impact assessments.

KEYWORDS: synthetic patient data, re-identification risk, privacy metrics, CPRD, data governance, differential privacy

DOI: 10.69554/LQOM5698

INTRODUCTION

In recent years, there has been much discussion about synthetic data with it being named as one of the emerging technologies in the healthcare data science space that is expected to become mainstream in two to five years, according to Gartner in their report, 'Hype Cycle for Healthcare Data, Analytics and AI 2023'.[1] Synthetic data was presented as a privacy enhancing technology (PET) to overcome challenges related to accessing real healthcare patient data because of privacy concerns. Guidance on PETs from the UK Information Commissioner's Office (ICO) recommended the use of synthetic data for generating non-personal data in situations where personal information could not be or did not need to be shared.[2]

More recently, however, questions have been asked about whether it is safe to assume that synthetic data is inherently private or, indeed, whether it could be considered personal data in some scenarios.[3,4] This paper will therefore provide an overview of synthetic data including types of synthetic data and different approaches to generating synthetic data that may have an impact on privacy in terms of reidentification. It will then discuss the various applications of synthetic patient data, with a focus on 'high-fidelity' synthetic data, before considering privacy aspects and debates around whether synthetic data could be considered personal data. The focus will be synthetic versions of tabular healthcare patient data and topics like synthetic imaging data or generated data yielded by large language models (LLMs) are outside the scope of this paper. The paper will include a case study from the UK Clinical Practice Research Datalink (CPRD) to illustrate how data custodians can assure themselves of the privacy of synthetic data and what checks they can put in place before data release to mitigate reidentification risks.

SYNTHETIC DATA: TYPES AND APPROACHES TO GENERATION

Synthetic data are artificial data crafted to reflect statistical properties, patterns and relationships in real-world datasets. They are generated or simulated rather than directly collected from authentic sources. The common approaches[5] to generating synthetic data include statistical-based, noise-based and machine-learning-based generation methods. Privacy and fidelity considerations will determine the choice of synthetic data generation approach.

Statistical-based methods involve replicating statistical properties of the original data, such as mean, variance and distribution. For instance, knowledge of the average blood pressure values and their distributional characteristics can be used to generate blood pressure readings. These methods rely on mathematical models to create synthetic datasets that closely resemble the statistical characteristics of the real data. A key limitation is that while each synthetic data field will have the statistical properties of real data, the complex relationship between data fields will be difficult to capture (eg the synthetic blood pressure readings may not reflect clinical expectations given other synthetically generated clinical characteristics when using this approach).

Noise-based methods, on the other hand, introduce random variations or 'noise' to the original data to create synthetic datasets. This approach involves perturbation of some of the data fields in real data in different ways. Perturbation could be accomplished by substitution of real values with other realistic values or random shuffling of data values within a particular data field or application of a random numeric variance (eg adding or subtracting 10 per cent to/from all data values in a field in such a way that the data distribution is preserved). These methods produce synthetic data with a slightly higher privacy risk compared to other synthetic generation methods, as they are only partially synthetic.

Machine-learning-based generation methods leverage algorithms and models to learn and replicate the underlying patterns and relationships present in real data. This category includes techniques such as generative adversarial networks (GANs) and variational autoencoders (VAEs), which can generate synthetic data with complex structures and dependencies. The learned data patterns are then used as an input for the synthetic data generator to yield synthetic data. Synthetic data generated using machine-learning techniques are better able to capture complex relationships between various data fields. In principle, like statistical approaches, these methods can yield synthetic data that presents a low reidentification risk as they are fully synthetic even though they closely resemble the original data. The main concern

regarding reidentification risk arises from the possibility of inferring information about individuals in the real-world data from the synthetic records, which underlines the importance of proactively preventing inference following an adversarial attack as illustrated in the CPRD case study later.

A recent development of the approach is towards a combination of machine-learning and noise-based approaches,[6] such as differential privacy,[7] to strike a balance between privacy and fidelity. For instance, using a privatised data generator[8] can enhance privacy protection while preserving fidelity. Machine learning methods and differential privacy methods are sometimes used in combination, leveraging the strengths of both approaches to achieve the best results. This synergy between machine learning and differential privacy can result in the generation of synthetic data with improved privacy guarantees and fidelity.

APPLICATIONS OF HIGH-FIDELITY SYNTHETIC DATA

The quality of synthetic data hinges on its 'fidelity', which determines how effectively it replicates the relevant features of the original data. High-fidelity synthetic data refers to synthetic data capable of capturing the intricate interrelationships that exist between various data fields, replicating the complex patterns observed in real data. In the context of healthcare data, a high-fidelity synthetic dataset would not only reflect the statistical properties (ie summary statistics) of the real data it is based on, it would also be clinically indistinguishable from the real data when reviewed by human clinical experts. A related concept when describing synthetic data, is its utility, which is a measure of how usable it is. A low or medium fidelity synthetic dataset could still have high utility depending on the purpose it is used for.

Many data protection professionals are now familiar with the use of synthetic data as a PET and both low-medium and high-fidelity synthetic datasets can be used in this context. Low and medium fidelity synthetic data is best suited to understanding the structure of the real data for writing programming code which will subsequently be used to query real data within a trusted research environment (TRE). High-fidelity synthetic data, however, could be used as a replacement for real data to train or validate machine learning algorithms.[9] The recently concluded Veterans Cardiac Health and AI Model Predictions (V-CHAMPs) Challenge, which was co-hosted by the Veterans Health Association (VHA) Innovation Ecosystem, the FDA Digital Health Center of Excellence, precisionFDA (part of the FDA's Office of Digital Transformation) and the Medicines and Healthcare products Regulatory Agency (MHRA), demonstrated the potential of synthetic data for the initial training and testing of AI algorithms.[10] In this case, research teams across the globe were invited to initially use a high-fidelity synthetic data version of a VHA dataset on heart failure to develop an AI model that could predict the risk of heart failure outcomes for the purpose of early clinical intervention. Shortlisted teams were then invited to work with the VHA to refine and validate their AI models on real data. It was found that the performance of the AI models trained on synthetic data, translated well to the real data.

High-fidelity synthetic data has a value beyond a PET and even when it is possible to access real data within a robust governance framework and privacy measures, high-fidelity synthetic data may be particularly useful for addressing limitations in real data such as missing data values[11] and biases due to underrepresentation of specific population subgroups.[12] Real-world data such as routinely collected electronic healthcare record (EHR) data may lack representation for specific patient groups, leading to biased results that could lead to unfair policies. It is possible to selectively boost underrepresented groups in the real

data by using synthetic data. These are examples of data augmentation whereby synthetic data is used to enhance real data.[13]

Finally, high-fidelity synthetic data generation methodologies could be applied to generate virtual patient cohorts for clinical trials to assess the safety, efficacy and benefits of medical interventions. This builds on the concept of *in silico* trials, which use computer simulation methods to demonstrate the efficacy and safety of medical products compared with traditional experimental (*in vitro* or *in vivo*) approaches of evidence generation. *In silico* trials may use historical or contemporary data from other clinical trials or real-world data sources to create virtual patient cohorts. Traditional methods can be time-consuming and costly, impeding medical innovation. More importantly, virtual patient cohorts can be used to simulate the effectiveness and safety of interventions in population subgroups, like pregnant women, children, the very elderly and immunocompromised, that are not usually included in clinical trials because of ethical or safety concerns.[14]

IS SYNTHETIC DATA PERSONAL DATA?

Not all the applications and benefits of synthetic patient data are contingent on its status as personal or non-personal data. However, clarity about the regulatory status of synthetic datasets and models is important because it has significant implications for the responsibilities of those developing and using the data. There is currently unhelpful ambiguity: despite evidence of privacy and reidentification risks that may arise from some synthetic data generation approaches,[15,16] high profile messaging in the form of the recently adopted EU AI Act crudely groups synthetic data with anonymised or 'other non-personal data' in several of its provisions.[17]

As yet, there is no court judgment or opinion on the key regulatory question of whether, or in what circumstances, synthetic data fall within the scope of 'personal data'.[18,19] However, data protection authorities across Europe have begun to consider synthetic data in discussion documents and emerging guidance, primarily as part of the assessment of privacy enhancing technologies and AI processing, and a consistent approach has begun to crystalise.[20] The data protection authorities are not approaching synthetic data as presumptively 'non personal data'. Instead, they are adopting what could be termed an 'orthodox' approach to synthetic data as a novel (privacy enhancing) technology. This begins with the position that data being processed is 'personal data' and a different conclusion will only be reached if data protection authorities are provided with a high degree of confidence that threats of reidentification are minimal and well safeguarded.[21]

Using patient data to generate synthetic data will therefore require a multifaceted assessment of whether the synthetic data model or its output data, in combination with other available sources, could identify a living individual. Data controllers will need to consider the 'means reasonably likely to be used'[22] to identify an individual, the technical and organisational privacy measures in places and the dataset's current and future processing environment and purposes in any given assessment. Of course, the prospect of identifying an individual does not have to be impossible for data to fall outside the scope of personal data but the standard set by the Court of Justice of the European Union (CJEU) is high, ruling that data would not be personal data if 'the risk appears in reality to be insignificant' or if it was practically impossible to identify an individual.[23]

While this approach to synthetic data maintains high standards of privacy and data protection, it should be acknowledged that it is not without cost and challenge. It may require resource and time-intensive auditing and adjustments to the data environment,

potentially leading to a reduction in synthetic data generation, higher access fees and limited accessibility for health researchers or developers. As the authors of this paper have already discussed, there are a multiplicity of forms that synthetic data may take, and it could be argued that a more proportionate regulatory approach should be adopted following careful deliberation by technical experts, regulators and policy makers. For example, it is possible that shifting the presumption that synthetic data are personal data for some forms of fully synthetic data generation (with controls like the removal of outliers for example) might be appropriate and defensible in certain circumstances.

However, regulatory consideration of synthetic data is still at an early stage and for the time being, synthetic data developers and users should continue to follow best practice[24] in relation to data protection impact assessments and anonymisation in assessing the identifiability and other data protection risks arising from the generation and processing of synthetic data. A key component of this is the consideration and adoption of privacy preservation techniques to safeguard synthetic data outputs.

PRIVACY PRESERVATION APPROACHES

For privacy preservation in synthetic data, the authors propose categorising techniques into two main groups: differential privacy, which seeks to give a provable privacy guarantee in the worst case, and statistical approaches that seek to quantify the level of identifiability within a dataset, thus attempting to rule out several kinds of risk of disclosure.

Differential privacy

Since its inception in 2006, the concept of differential privacy[25] has evolved from a theoretical framework into a practical tool with significant real-world applications. Initially introduced by Cynthia Dwork and colleagues, it provided a rigorous approach for privacy-preserving data analysis. Unlike traditional approaches that aim to protect the final analysis output of a computation, differential privacy aims to protect the privacy of individuals' data throughout the data analysis process. Over the years, researchers and practitioners have developed practical techniques to implement differential privacy, leading to broader adoption across academia and industry. Standardisation efforts and regulatory frameworks, such as those by NIST,[26] have further solidified its role as a key PET. Recent advancements have expanded its applications and improved its utility, with ongoing research focusing on enhancing privacy guarantees while maintaining data utility,[27] especially in the context of emerging technologies like machine learning[28] and federated learning.[29]

The key idea behind differential privacy is to add controlled noise or randomness to the data before any analysis is performed. This noise ensures that the presence or absence of any individual's data has minimal impact on the results of the analysis. By doing so, differential privacy prevents adversaries from determining whether a specific individual's data was included in the analysis, thereby preserving the privacy of individuals in the dataset.

Differential privacy provides robust guarantees against adversaries with significant computational power and access to external datasets. It ensures that even if an adversary has access to auxiliary information or external datasets, they cannot reliably infer whether a specific individual's data was used in the analysis. This makes it significantly harder for adversaries to conduct privacy attacks or identify individuals based on the released data.

While differential privacy offers strong worst-case guarantees in terms of privacy protection, it may not be universally applicable due to its stringent requirements.

For instance, implementing differential privacy can sometimes lead to a trade-off between privacy and data utility, as adding noise to the data may affect the accuracy or usefulness of the analysis results. Additionally, differential privacy requires careful parameter tuning and consideration of the specific privacy requirements of the dataset and analysis task. While differential privacy is typically applied to real datasets, it can also be considered as part of the synthetic data generation process though the same caveats about the trade-off between privacy and fidelity apply. As discussed previously, this may not matter for some synthetic data applications as loss of fidelity does not always translate into a loss of utility. It could, however, adversely impact applications needing high-fidelity synthetic datasets and a high degree of accuracy.

Expert determination: Quantifying the level of identifiability

These approaches rely on first quantifying the level of identifiability within a tabular dataset, and then, if this exceeds the risk threshold considered acceptable by a data controller, putting in place measures such as aggregation (for the purpose of generalisation) or suppression of certain data categories, until an acceptable threshold is achieved. This approach relies on privacy metrics to assess the output (ie the synthetic dataset generated) and evaluate privacy preservation based on factors such as k-anonymity,[30] l-diversity[31] and t-closeness.[32] They share a common working mechanism of ensuring that individual records in a dataset are grouped together in a way that makes them indistinguishable or less identifiable.

K-anonymity requires that each record is part of a group of at least k similar records (ie hiding in a crowd). However, if the adversary knows that a given individual is in a dataset along with some of their attributes, and the dataset does not hold sufficient diversity, they could infer other sensitive attributes of this individual that were previously unknown to them. An example scenario is where the adversary knows the age, sex and region of an individual, they could infer that this individual has diabetes if all other individuals in that dataset sharing the same age, sex and region, have diabetes. Therefore, the adversary would not need to identify which record in the dataset relates to this individual to infer their diabetic status (homogenous pattern attack or background knowledge attack). This risk can be minimised via l-diversity by requiring that each group contains at least l different values for sensitive attributes. Another useful metric to quantify privacy risks is t-closeness, which focuses on minimising the statistical distance between the distribution of sensitive attributes in the dataset and the distribution in the broader population. These metrics can be used to guide data custodians in setting acceptable parameter values that reduce the risk of reidentification while preserving the utility of the data for analysis purposes.

Many data custodians choose a k-anonymisation level of 10, ie at least 10 individuals in the dataset share a sensitive characteristic. The l-diversity metric is usually less than or equal to the chosen value of k but more than 1. A suggested t-closeness value may be 0.1, ie the distribution of a sensitive characteristic within each stratum of quasi-identifiers (eg a stratum based on age group and sex) should be 10 per cent of the distribution of that characteristic within the entire dataset. There is no magic number, and ultimately, the choice of anonymisation parameter value is based on the data custodian's risk appetite and the intended purpose of the data. For instance, the European Medicines Agency recommend a reidentification risk threshold[33] of 0.09 for the publication of clinical data for medicinal products, ie that the chance that someone in a group with similar information could be identified from anonymised data is not more than

9 per cent. This can be achieved through adjusting the k value, l value and t value to calculate the final reidentification risk value. The authors of this paper suggest that the choice of anonymisation parameter values is made by a team of individuals representing expertise in information governance, the data context and intended applications of the data.

CASE STUDY: CPRD SYNTHETIC DATA

This section focuses on a case study concerning the implementation of synthetic patient data within the CPRD, which is a UK government real-world data research service hosted within the MHRA, placing particular emphasis on privacy preservation approaches. The case study delves into the reidentification risk assessment and privacy considerations involved in generating synthetic datasets, highlighting the challenges and strategies for safeguarding privacy in healthcare research.

ABOUT THE CPRD DATA

CPRD makes anonymised primary care patient data available for public health research and researchers need to go through a data access application process and sign a data licence agreement before they can access the data.[34] More recently, CPRD has made available two high-fidelity and two medium-fidelity synthetic datasets based on CPRD data.[35] The high-fidelity synthetic datasets are derived from CPRD primary care data so while they can be used instead of real patient data for statistical analysis and machine learning applications, they cannot be used to understand the base structure of the CPRD data. The medium-fidelity synthetic datasets on the other hand, resemble the real-world CPRD data with respect to the data types, data values, data formats, data structure and table relationships. While the medium-fidelity datasets cannot be used for statistical or machine learning analysis, they can be used for multiple other purposes including as a sample dataset to understand the structure and utility of the anonymised databases, to use as a data management teaching/training resource, to develop/validate/test analytics tools for use with CPRD data (eg a bespoke cohort selection tool) or to develop machine-learning workflows that can be applied to anonymised real-world CPRD data.

PRIVACY PRESERVATION AND RISK ASSESSMENT

CPRD sought advice from the ICO Innovation Hub,[36] when the initial development of synthetic datasets was considered. CPRD was advised to undertake a data protection impact assessment (DPIA) risk assessment before releasing the synthetic data and to consider releasing the data under a data access agreement (DAA). The CPRD Information Governance team conducted a DPIA that involved them requesting, from the technical team, a full description of the methods used to generate the synthetic datasets, including the implementation of privacy preservation approaches. The initial synthetic data generation used CPRD primary-care data that had already been anonymised in line with the ICO anonymisation code of conduct, as the basis. This meant that the starting confidentiality risks were remote. In addition, due to concerns raised by the lay representative on the MHRA's Synthetic Data Project Steering Group, privacy measures were integrated throughout the synthetic data generation process as outlined in Wang *et al*.[37]

In the case of CPRD, the synthetic data generation used a Bayesian Network (BN) approach (that could be categorised as a machine learning approach), whereby, a BN analysis was first used to learn the patterns in the anonymised real-world CPRD data (the ground truth data). These patterns, rather than real data, were fed into the synthetic data generator to create 100 per cent artificial patient records. However, the

Steering Group noted that it was possible that a synthetic patient record could be created that coincidentally had the same features as a real record in the ground truth data. The Steering Group thus advised CPRD to remove any exact duplicates between the generated synthetic data and the ground truth CPRD data. Despite the anonymisation of CPRD primary-care data, theoretical reidentification risks prompted additional controls such as adding noise through perturbation techniques and removing rare conditions to obscure potential reidentification paths. CPRD opted not to use a differential privacy approach as the nature of the data and intended purpose meant that any privacy gains would be at the cost of unacceptable losses in data utility.

CPRD undertook a simulated adversary attack, where a malicious actor managed to obtain the full synthetic dataset and part of the ground truth dataset, to assess the risk of linking any synthetic data record to a real patient record in the ground truth data. This exercise found that the risk of doing so was remote. In the scenario tested, even if the adversary highlighted possible matches between synthetic data candidates and the ground truth data, it was not possible for them to assert with any certainty that these were true matches rather than just coincidental similarities. In fact, because the CPRD synthetic data generation method involved the removal of exact duplicates and rare patterns, the researcher playing the role of the adversary was unable to link any of the synthetic data records to the ground truth data and nor were they able to exploit methodological approaches like outlier detection and k-nearest neighbour to assess similarity.

ADDITIONAL REIDENTIFICATION RISK MITIGATION MEASURES

Despite rigorous obscuring, CPRD's information governance team advised the implementation of additional measures to mitigate even the extremely remote reidentification risks. CPRD was advised to manage the dataset similarly to anonymised CPRD real-world data, emphasising strict controls on access, data sharing and destruction. In addition, CPRD was advised against sharing the exact ground truth data extract used for the initial BN analysis to learn data patterns for the purpose of the synthetic data generation. Moreover, CPRD was advised to not share the full programming code for the synthetic data generation, the detailed simulated adversary attack report or the DPIA reports outside CPRD to prevent any malicious attempts at reverse engineering any real patient records from the synthetic data.

Additional mitigation strategies included new client review checks for applicants (albeit much lighter touch than those required for accessing real anonymised data), restricted dataset sharing and enforcing data destruction at the end of the licensing term. Furthermore, any subsequent dataset refreshes would require retiring the current version and implementing stringent data destruction protocols before accessing the new version. All these recommendations were accepted and implemented by CPRD to enable sharing of synthetic data while providing assurance to patients and data regulators that reidentification risks have been mitigated to the point of being exceedingly remote.

CONCLUSIONS

The potential applications of high-fidelity synthetic patient data in healthcare research are numerous. In this paper, the authors have explored the challenges and benefits associated with using synthetic data. The paper proposes a definition of synthetic data, outlines different approaches to synthetic data generation methods, and summarises the potential applications of high-fidelity synthetic data.

As synthetic data continues to gain traction in healthcare research, the question of whether it should be considered personal data remains a topic of debate. While regulatory frameworks and data protection authorities are beginning to address this issue, there remains ambiguity surrounding the classification of synthetic data and its implications for data governance.

Privacy preservation approaches, including differential privacy and statistical methods for quantifying identifiability, play a crucial role in safeguarding the privacy of individuals' data throughout the data analysis process. By implementing some or all of these techniques, data custodians can ensure that synthetic datasets maintain a high level of privacy protection while preserving their utility for research purposes.

The case study on synthetic healthcare patient data from CPRD illustrates how privacy preservation approaches, risk assessment and mitigation strategies can be implemented to enable the sharing of synthetic datasets in a way that balances privacy and utility, even amidst the current ambiguity. In the longer term, continued exploration is needed of privacy preservation approaches and regulatory considerations with the aim of further streamlining the governance around synthetic data so that their full potential can be harnessed for research and innovation.

The authors also urge the research community to study the cost-benefit relationship of synthetic data sharing, and appropriately weigh the risk of privacy disclosure against benefits of increased data use.

References

1. Gartner (24th July, 2023) 'Hype Cycle for Healthcare Data, Analytics and AI', available at https://www.gartner.com/en/documents/4557999 (accessed 21st March, 2024).
2. ICO (n.d.) 'Synthetic Data', available at https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-sharing/privacy-enhancing-technologies/what-pets-are-there/synthetic-data/ (accessed 13th March, 2024).
3. PHG Foundation (n.d.) 'Are Synthetic Health Data "Personal Data"?', available at https://www.phgfoundation.org/publications/reports/are-synthetic-health-data-personal-data/ (accessed 21st March, 2024).
4. Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., Cohen, S. N. and Weller, A. (2022) 'Synthetic Data — What, Why and How?', arXiv, available at https://doi.org/10.48550/arXiv.2205.03257 (accessed 21st March, 2024).
5. Wang, Z., Myles, P. and Tucker, A. (2021) 'Generating and Evaluating Cross-Sectional Synthetic Electronic Healthcare Data: Preserving Data Utility and Patient Privacy', *Computational Intelligence*, Vol. 37, No. 2, pp. 819–51. doi:10.1111/coin.12427.
6. Kopp, A. (18th February, 2021) 'Create Privacy-preserving Synthetic Data for Machine Learning with SmartNoise', Microsoft, available at https://cloudblogs.microsoft.com/opensource/2021/02/18/create-privacy-preserving-synthetic-data-for-machine-learning-with-smartnoise/ (accessed 21st March, 2024).
7. Dwork, C. (2006) 'Differential Privacy', *Automata, Languages and Programming*, Vol. 2006, pp. 1–12. doi:10.1007/11787006_1.
8. Kopp, ref 6 above.
9. Wang *et al.*, ref 5 above.
10. Precision FDA (2023) 'The Veterans Cardiac Health and and AI Model Predictions (V-CHAMPS) Challenge 2023', available at https://precision.fda.gov/challenges/31 (accessed 21st March, 2024).
11. Tucker, A., Wang, Z., Rotalinti, Y. and Myles, P. (2020) 'Generating High-fidelity Synthetic Patient Data for Assessing Machine Learning Healthcare Software', *npj Digital Medicine*, Vol. 3, No. 1, Article 147. doi:10.1038/s41746-020-00353-9.
12. Draghi, B., Wang, Z., Myles, P. and Tucker, A. (2024) 'Identifying and Handling Data Bias within Primary Healthcare Data Using Synthetic Data Generators', *Heliyon*, Vol. 10, No. 2, Article e24164. doi:10.1016/j.heliyon.2024.e24164.
13. Wang, Z., Draghi, B., Rotalinti, Y., Lunn, D. and Myles, P. (2024) 'High-fidelity Synthetic Data Applications for Data Augmentation', IntechOpen, available at https://www.intechopen.com/online-first/88698 (accessed 21st March, 2024).
14. Myles, P., Ordish, J. and Tucker, A. (2023) 'The Potential Synergies between Synthetic Data and In Silico Trials in Relation to Generating Representative Virtual Population Cohorts', *Progress in Biomedical Engineering*, Vol. 5, No. 1, Article 013001. doi:10.1088/2516-1091/acafbf.
15. Stadler, T., Oprisanu, B. and Troncoso, C. (1970) 'Synthetic Data — Anonymisation Groundhog Day', available at https://www.usenix.org/conference/usenixsecurity22/presentation/stadler (accessed 21st March, 2024).
16. Chen, D., Yu, N., Zhang, Y. and Fritz, M. (2020) 'Gan-leaks: A Taxonomy of Membership Inference

Attacks against Generative Models', Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security. doi:10.1145/3372297.3417238.
17. European Parliament legislative resolution of 13 March 2024 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts, available at https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.html (accessed 14th February, 2024).
18. Parliament and Council Regulation (EU) 2016/679 of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), Article 4(1).
19. ICO, ref 2 above.
20. *Ibid*.
21. *Ibid*.
22. GDPR, ref 18 above, Recital 26.
23. Case C-582/14 *Patrick Breyer v Bundesrepublik Deutschland* [2016] ECR I-769, para 49.
24. UK Anonymisation Network (n.d.), 'The ADF', available at: https://ukanon.net/framework/ (accessed 13th March, 2024).
25. Dwork, ref 7 above.
26. Near, J. P. and Darais, D. (2023) 'Guidelines for Evaluating Differential Privacy Guarantees', NIST, available at https://doi.org/10.6028/NIST.SP.800-226.ipd (accessed 1st May, 2024).
27. Wang, T., Ding, B., Xu, M., Huang, Z., Hong, C., Zhou, J., Li, N. and Jha, S. (2020) 'Improving Utility and Security of the Shuffler-based Differential Privacy', *Proceedings of the VLDB Endowment*, Vol. 13, No. 13, pp. 3545–58.
28. Ponomareva, N., Vassilvitskii, S., Xu, Z., McMahan, B., Kurakin, A. and Zhang, C. (4th August, 2023) 'How to DP-fy ML: A Practical Tutorial to Machine Learning with Differential Privacy', *Journal of Artificial Intelligence Research*, Vol. 77, pp. 1113–201.
29. Zhang, X., Kang, Y., Chen, K., Fan, L. and Yang, Q. (2023) 'Trading Off Privacy, Utility and Efficiency in Federated Learning', *ACM Transactions on Intelligent Systems and Technology*, Vol. 14, No. 6, Article 98.
30. Sweeney, L. (2002) 'K-anonymity: A Model for Protecting Privacy', *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol. 10, No. 05, pp. 557–70. doi:10.1142/s0218488502001648.
31. Machanavajjhala, A., Gehrke, J., Kifer, D. and Venkitasubramaniam, M. (2006) 'L-diversity: Privacy beyond K-anonymity', 22nd International Conference on Data Engineering (ICDE'06). doi:10.1109/icde.2006.
32. Li, N., Li, T. and Venkatasubramanian, S. (2007) 'T-closeness: Privacy beyond K-anonymity and L-diversity', 2007 IEEE 23rd International Conference on Data Engineering. doi:10.1109/icde.2007.367856.
33. European Medicines Agency (2016) 'External Guidance on the Implementation of the European Medicines Agency Policy on the Publication of Clinical Data for Medicinal Products for Human Use', available at https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/external-guidance-implementation-european-medicines-agency-policy-publication-clinical-data-medicinal-products-human-use-first-version_en.pdf (accessed 21st March, 2024).
34. CPRD (31st October, 2023) 'Data Access', available at https://cprd.com/data-access (accessed 21st March, 2024).
35. CPRD (12th March, 2024) 'Synthetic Data', available at https://cprd.com/synthetic-data (accessed 21st March, 2024).
36. ICO (n.d.) ICO 'Innovation Services', available at https://ico.org.uk/about-the-ico/what-we-do/ico-innovation-services/#ih (accessed 21st March, 2024).
37. Wang *et al.*, ref 5 above.